Covering numbers and its application on deep neural networks
Chapter 20-chapter 21

Presenter: Xianli Zeng

May 13, 2021

Outline

Motivation

Let x_i are IID samples from distribution $\mathbb{P}$. Its empirical distribution is denoted by $\mathbb{P}_n$. In statistical learning, we are interested in

$$||\mathbb{P}_n - \mathbb{P}||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]|.$$

- When $|\mathcal{F}|$ is finite, the bound is easily derived by concentration inequalities.
- When $|\mathcal{F}|$ is infinite, in previous chapter, we have shown that the above quantity can be upper and lower bounded by the Rademacher complexity of $\mathcal{F}$.

A straightforward idea

- To provide a uniform bound for a set U with infinite number of elements is difficult. In converse, we first consider a set V with finite number of elements.
- How to choose V: we want that for any $u \in U$, there is a $v \in V$ such that u and v share similar properties.

To be specific, we want to bound $sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]|$. The first step is, for any ϵ, we find N_ϵ elements $f_1, ..., f_{N_\epsilon}$ of $\mathcal{F}$ satisfies, for any $f \in \mathcal{F}$, there exists a i such that

$$\|f - f_i\|_\infty \leq \epsilon.$$

Denote $\phi(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]$. Then,

$$\sup_{f\in\mathcal{F}} |\phi_f| = \sup_{f\in\mathcal{F}} |\phi_f - \phi_{f_i} + \phi_{f_i}| \leq 2\epsilon + \sup_i |\phi_{f_i}|.$$

Covering number: definition

Definition

Given a set U, scale ϵ, norm $\|\cdot\|$, $V \subset U$ is a (proper) ϵ-cover when

$$\sup_{u \in U} \inf_{v \in V} \|u - v\| \leq \epsilon.$$

Let $\mathcal{N}(U, \epsilon, \|\cdot\|)$ denote the covering number: the cardinality of the smallest ϵ-cover.

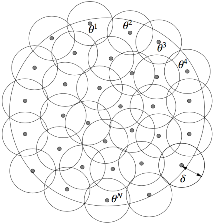

Figure: illustration of covering number

Examples (Covering numbers of unit interval)

Let us begin with a simple example. Consider the interval $[-1, 1]$ in $\mathbb{R}$, equipped with the norm $\|\theta\| = |\theta|$.
Suppose that we divide the interval $[-1, 1]$ into $L := \lfloor \frac{1}{\epsilon} \rfloor + 1$, sub-intervals, centered at the points

$$\theta^i = -1 + 2(i-1)\epsilon \qquad \text{for} \qquad i \in [L] : \{1, 2, ..., L\},$$

and each of length at most 2ϵ. By construction, for any point $\theta \in [0, 1]$, there is some $j \in [L]$ such that

$$|\theta^j - \theta| \leq \epsilon,$$

which shows that

$$\mathcal{N}([-1, 1], \epsilon, \cdot|) \leq \lfloor \frac{1}{\epsilon} \rfloor + 1.$$

Basic Rademacher-covering relationship: one step discretization bounds

Covering numbers and Rademacher complexities are in some usual settings nearly tight with each other, though in these lectures we will only produce a way to upper bound Rademacher complexity with covering numbers.

Theorem

Given $U \subset \mathbb{R}^n$,

$$URad(U) \leq \inf_{\alpha > 0} \left(\alpha\sqrt{n} + (\sup_{u \in U} \|u\|_2^2)\sqrt{2 \ln \mathcal{N}(U, \alpha, \|\cdot\|_2)} \right).$$

proof

Proof.

Let $\alpha > 0$ be arbitrary, and suppose $\mathcal{N}(U, \alpha, \|\cdot\|_2) < \infty$ (otherwise bound holds trivially). Let V denote a minimal cover, for any $u \in U$, denote $V(u)$ the closest element in V.

$$
\begin{aligned}
\mathrm{URad}(U) &= \mathbb{E} \sup_{u \in U} \langle \epsilon, u \rangle \\
&= \mathbb{E} \sup_{u \in U} \langle \epsilon, u - V(u) + V(u) \rangle \\
&= \mathbb{E} \sup_{u \in U} (\langle \epsilon, V(u) \rangle + \langle \epsilon, u - V(u) \rangle) \\
&= \mathbb{E} \sup_{u \in U} (\langle \epsilon, V(u) \rangle + \|\epsilon\|_2 \|u - V(u)\|_2 \\
&= \mathrm{URad}(V) + \alpha\sqrt{n} \\
&= \sup_{v \in V} (\|v\|_2) \sqrt{2 \ln |V|} + \alpha\sqrt{n} \\
&= \sup_{u \in U} (\|u\|_2) \sqrt{2 \ln \mathcal{N}(U, \alpha, \|\cdot\|_2)} + \alpha\sqrt{n}
\end{aligned}
$$

and the bound follows since $\alpha > 0$ was arbitrary. □

Second Rademacher-covering relationship: Dudley's entropy integral

Theorem

Let $U \subseteq [-1,+1]^n$ *be given with* $0 \in U$*.*

$$\begin{aligned} URad(U) &\leq \inf_{N \in \mathbb{Z}_{\geq 1}} \left(n2^{1-N} + 6\sqrt{n} \sum_{i=1}^{N} 2^{-i} \sqrt{\ln \mathcal{N}(U, 2^{-i}\sqrt{n}, \|\cdot\|^2)} \right) \\ &\leq \inf_{\alpha > 0} \left(4\alpha\sqrt{n} + 12 \int_{\alpha}^{\sqrt{n}/2} \sqrt{\ln \mathcal{N}(U, \beta, \|\cdot\|^2)} d\beta \right). \end{aligned}$$

proof

Proof.

We'll do the discrete sum first. The integral follows by relating an integral to its Riemann sum.

- Let $N \geq 1$ be arbitrary.
- For $i \in \{1, ..., N\}$, define scales $\alpha_i := \sqrt{n}2^{1-i}$.
- Define cover $V_1 := \{0\}$; since $U \subseteq [-1, +1]^n$, this is a minimal cover at scale $\alpha = \sqrt{n}$.
- Let V_i for $i \in \{2, ..., N\}$ denote any minimal cover at scale α_i, meaning $|V_i| = \mathcal{N}(U, \alpha_i, \|\cdot\|_2)$.

$\square$

Proof.

Since

$$u = (u - V_N(u)) + \sum_{i=1}^{N-1} (V_{i+1}(u) - V_i(u)) + V_1(u),$$

$$\begin{aligned}
\mathrm{URad}(U) &= \mathbb{E} \sup_{u \in U} \langle \epsilon, u \rangle \\
&= \mathbb{E} \sup_{u \in U} \left(\langle \epsilon, (u - V_N(u)) + \sum_{i=1}^{N-1} (V_{i+1}(u) - V_i(u)) + V_1(u) \rangle \right) \\
&= \mathbb{E} \sup_{u \in U} \langle \epsilon, u - V_N(u) \rangle + \sum_{i=1}^{N-1} \mathbb{E} \sup_{u \in U} \langle \epsilon, V_{i+1}(u) - V_i(u) \rangle + \mathbb{E} \sup_{u \in U} \langle \epsilon, V_1(u) \rangle
\end{aligned}$$

□

Proof.

Combining these bounds,

$$\mathrm{URad}(U) \leq n2^{1-N} + 0 + 6\sqrt{n}\sum_{i=1}^{N} 2^{-i}\sqrt{\ln \mathcal{N}(U, 2^{-i}\sqrt{n}, \|\cdot\|^2)}.$$

$N \geq 1$ was arbitrary, so applying $\inf_{N\geq 1}$ gives the first bound.
For the second bound, as $\mathcal{N}(U, \beta, \|\cdot\|^2)$ is nonincreasing in β, the integral upper bounds the Riemann sum:

$$\begin{aligned}\mathrm{URad}(U) &\leq n2^{1-N} + 12\sum_{i=1}^{N}(\alpha_{i+1} - \alpha_{i+2})\sqrt{\ln \mathcal{N}(U, 2^{-i}\sqrt{n}, \|\cdot\|^2)} \\ &\leq \sqrt{n}\alpha_N + 12\int_{\alpha_{N+1}}^{\alpha_2} \sqrt{\ln \mathcal{N}(U, \beta, \|\cdot\|^2)}d\beta.\end{aligned}$$

Tofinish,pick $\alpha > 0$ and N with

$$\alpha_{N+1} \geq \alpha > \alpha_{N+2}.$$

$\square$

Outline

Bounds of covering numbers

We will give bounds of covering numbers of two different groups of functions.

- The first will be for arbitrary Lipschitz functions, and will be horifically loose (exponential in dimension).
- The second will be the tightest known bound for ReLU networks.

Covering number of Lipschitz functions

Theorem

Let data $S = (x_1, \ldots, x_n)$ be given with $R := \max_{i,j} \|x_i - x_j\|_\infty$. Let $\mathcal{F}$ denote all ρ-Lipschitz functions from $[-R, +R]^d \to [-B, +B]$ (where Lipschitz is measured wrt $\|\cdot\|_\infty$). Then the improper covering number $\widetilde{\mathcal{N}}$ satisfies

$$\ln \widetilde{\mathcal{N}}(\mathcal{F}, \epsilon, \|\cdot\|_u) \leq \max\left\{0, \left\lceil \frac{4\rho(R+\epsilon)}{\epsilon} \right\rceil \ln \left\lceil \frac{2B}{\epsilon} \right\rceil \right\}.$$

Covering number of Lipschitz functions

Proof.

- Suppose $B > \epsilon$, otherwise can use the trivial cover $\{x \to 0\}$.
- Subdivide $[-R-\epsilon, +R+\epsilon]^d$ into $(\frac{4(R+\epsilon)\rho}{\epsilon})^d$ cubes of side length $\frac{\epsilon}{2\rho}$; call this U.
- Subdivide $[-B, +B]$ into intervals of length ϵ, thus $2B/\epsilon$ elements; call this V .
- Our candidate cover G is the set of all piecewise constant maps from $[-R-\epsilon, +R+\epsilon]^d$ to $[-B, +B]$ discretized according to U and V , meaning

$$|\mathcal{G}| \leq \left\lceil \frac{2B}{\epsilon} \right\rceil^{\left\lceil \frac{4(R+\epsilon)\rho}{\epsilon} \right\rceil^d}$$

$\square$

Covering number of Lipschitz functions

Proof.

To show this is an improper cover, given $f \in \mathcal{F}$, choose $g \in \mathcal{G}4$ by proceeding over each $C \in U$, and assigning $g_{|C} \in V$ to be the closest element to $f(x_C)$, where x_C is the midpoint of C. Then,

$$\begin{aligned}
\|f-g\|_\infty &= \sup_{C\in U}\sup_{x\in C}|f(x)-g(x)| \\
&\leq \sup_{C\in U}\sup_{x\in C}(|f(x)-f(x_C)|+|f(x_C)-g(x)|) \\
&\leq \sup_{C\in U}\sup_{x\in C}(\rho\|x-X_C\|_\infty+\frac{\epsilon}{2}) \\
&\leq \sup_{C\in U}\sup_{x\in C}(\rho\frac{\epsilon}{4\rho}+\frac{\epsilon}{2}) \leq \epsilon.
\end{aligned}$$

$\square$

covering number bound based on partition

We now introduce the covering number for deep neural networks.

Theorem

Fix Relu activations σ and data $X \in \mathbb{R}^{n\times d}$, define

$$\mathcal{F}_n := \{f = W_L\sigma_{L-1}(\cdots\sigma_1(W_1X^\top)\cdots) : \|f\|_\infty \le R, \|W_i\|_{\infty,\infty} \le k, \\ \|b\|_\infty \le k, \|W_i\|_0 + \|b_i\|_0 \le S\},$$

and all matrix dimensions are at most m. Then

$$\mathcal{N}(\delta, \mathcal{F}_n, \|\cdot\|_\infty) \le (Lm^2)^S(\frac{2k}{h})^S \le (\frac{2L^2\|X\|_\infty k^L m^{L+2}}{\delta})^S,$$

Part 1 Construct a covering

Proof.

Since each weight parameter in the network is bounded by a constant k, we construct a covering by partition the range of each weight parameter into a uniform grid. Consider $f, f' \in \mathcal{F}(R, k, L, p, S)$ with each weight parameter differing at most h, i.e. $\|W_i - W_i'\|_{\infty,\infty} \leq h$ and $\|b_i - b_i'\|_\infty \leq h$. Denote

$$A_L = \|f - f'\|_\infty = \|W_L\sigma(W_{L-1}\cdots\sigma(W_1X)\cdots) - W_L'\sigma(W_{L-1}'\cdots\sigma(W_1'X)\cdots)\|_\infty,$$

By an induction on the number of layers in the network, we show that the norm of the difference $\|f - f'\|_\infty$ scales as $\square$

Part 2 Calculate δ

Proof.

$$
\begin{aligned}
\|f - f'\|_\infty &= A_L = \|W_L\sigma(W_{L-1}\cdots\sigma(W_1X)\cdots) - W_L'\sigma(W_{L-1}'\cdots\sigma(W_1'X)\cdots)\|_\infty \\
&\le \|W_L - W_L'\|_1\|W_{L-1}\cdots\sigma(W_1X)\cdots\|_\infty + \|W_L\|_1 A_{L-1} \\
&\le hm\|W_{L-1}\cdots\sigma(W_1X)\cdots\|_\infty + kmA_{L-1} \\
&\le hk^{L-1}m^L\|X\|_\infty + kmA_{L-1} \\
&\le hk^{L-1}m^L\|X\|_\infty + km(hk^{L-2}m^L\|X\|_\infty + kmA_{L-2}) \\
&= 2hk^{L-1}m^L\|X\|_\infty + k^2m^2A_{L-2} \\
&\le (L-1)hk^{L-1}m^L\|X\|_\infty + k^{L-1}m^{L-1}A_1 \\
&\le (L-1)hk^{L-1}m^L\|X\|_\infty + hk^{L-1}m^L\|X\|_\infty \\
&= hLk^{L-1}m^L\|X\|_\infty.
\end{aligned}
$$

$\square$

Part 3 Calculate covering number

Proof.

As a result, to achieve a δ-covering, it suffices to choose h such that $Lhk^{L-1}m^L\|X\|_\infty = \delta$. Moreover, there are $C^S_{Lm^2} \leq (Lm^2)^S$ different choices of S non-zero entries out of Lm^2 weight parameters. Therefore, the covering number is bounded by

$$\mathcal{N}(\delta, \mathcal{F}_n, \|\cdot\|_\infty) \leq (Lm^2)^S(\frac{2k}{h})^S \leq (\frac{2L^2\|X\|_\infty k^L m^{L+2}}{\delta})^S,$$

$\square$

“Spectrally-normalized” covering number bound

Theorem

Fix multivariate activations $(\sigma_i)_{i=1}^{L}$ *with* $\|\sigma\|_{Lip} =: \rho_i$ *and* $\sigma_i(0) = 0$, *and data* $X \in \mathbb{R}^{n\times d}$, *and define*

$$\mathcal{F}_n := \left\{ \sigma_L(W_L \sigma_{L-1} \cdots \sigma_1(W_1 X^\top) \cdots) : \|W_i^\top\|_2 \le s_i, \|W_i^\top\|_{2,1} \le b_i \right\},$$

and all matrix dimensions are at most m. *Then*

$$\ln \mathcal{N}(\mathcal{F}_n, \epsilon, \|\cdot\|_F) \le \frac{\|X\|_F^2 \Pi_{j=1}^L \rho_j^2 s_j^2}{\epsilon^2} \Big(\sum_{i=1}^{L} (\frac{b_i}{s_i})^{2/3}\Big)^3 \ln(2m^2).$$

Remark

Remark

Applying Dudley, we have

$$URad(\mathcal{F}_n) = \tilde{O}\left(\|X\|_F(\Pi_{j=1}^L \rho_j s_j)\left(\sum_{i=1}^L (\frac{b_i}{s_i})^{2/3}\right)^{3/2}\right).$$

Let's compare to our best "layer peeling" proof from before, which had $\Pi_i\|W_i\|_F \leq m^{L/2}\Pi_i\|W_i\|_2$. *That proof assumed* $\rho_i = 1$, *so the comparison boils down to*

$$m^{L/2}\Pi_i\|W_i\|_2 \qquad \text{and} \qquad \left[\sum_i \left(\frac{\|W_i^\top\|_{2,1}^{2/3}}{\|W_i\|_2^{2/3}}\right)\right]^{3/2}(\Pi_i\|W_i\|_2)$$

where $L \leq \sum_i \left(\frac{\|W_i^\top\|_{2,1}^{2/3}}{\|W_i\|_2^{2/3}}\right) \leq Lm^{2/3}$. *So the bound is better but still leaves a lot to be desired and is loose in practice.*

Two lemmas in proof

The first step of the proof is a covering number for individual layers,

Lemma

$$\ln \mathcal{N}(\{WX^\top : X \in \mathbb{R}^{m\times d}, \|W^\top\|_{2,1} \leq b\}, \epsilon, \|\cdot\|_F) \leq \left\lceil \frac{\|X\|_F^2 b^2}{\epsilon^2} \right\rceil \ln(2dm).$$

With the covering number for individual layers, we have the following covering number bound for the whole network,

Lemma

Let $\mathcal{F}_n$ be the same image vectors as in the theorem, and let per-layer tolerances $(\epsilon_1, ..., \epsilon_L)$ be given. then

$$\ln \mathcal{N}(\mathcal{F}_n, \sum_{j=1}^{L} \rho_j \epsilon_j \Pi_{k=j+1}^{L} \rho_k s_k, \|\cdot\|_F) \leq \sum_{i=1}^{L} \left\lceil \frac{\|X\|_F^2 b_i^2 \Pi_{j<i} \rho_j^2 s_j^2}{\epsilon_i}^2 \right\rceil \ln(2m^2).$$

Proof of the theorem

Proof.

We prove the theorem by solving a Lagrangian (minimize cover size subject to total error $\leq \epsilon$), choose

$$\epsilon_i = \frac{\alpha_i \epsilon}{\rho_i \Pi_{j>i} \rho_j s_j}, \qquad \alpha_i := \frac{1}{\beta}\left(\frac{b_i}{s_i}\right)^{2/3}, \qquad \beta := \sum_{i=1}^{L}\left(\frac{b_i}{s_i}\right)^{2/3}.$$

Invoking the induction lemma with these choices, the resulting cover error is

$$\sum_{i=1}^{L} \epsilon_i \rho_i \Pi_{j>i} \rho_j s_j = \epsilon \sum_{j=1}^{L} \alpha_i = \epsilon.$$

and the main term of the cardinality (ignoring $\ln(2m^2)$) satisfies

$$\sum_{i=1}^{L} \frac{\|X\|_F^2 b_i^2 \Pi_{j<i} \rho_j^2 s_j^2}{\epsilon_i^2} = \frac{\|X\|_F^2}{\epsilon^2} \sum_{i=1}^{L} \frac{b_i^2 \Pi_{j=1}^{L} \rho_j^2 s_j^2}{\alpha_i^2 s_i^2}$$

$$= \frac{\|X\|_F^2 \Pi_{j=1}^{L} \rho_j^2 s_j^2}{\epsilon^2} \sum_{i=1}^{L} \frac{\beta^2 b_i^{2/3}}{s_i^{2/3}} = \frac{\|X\|_F^2 \Pi_{j=1}^{L} \rho_j^2 s_j^2}{\epsilon^2} \left(\sum_{i=1}^{L}\left(\frac{b_i}{s_i}\right)^{2/3}\right)^3.$$

$\square$

Proof of the Lemma 1

Lemma

$$\ln \mathcal{N}(\{WX^\top : X \in \mathbb{R}^{m\times d}, \|W^\top\|_{2,1} \le b\}, \epsilon, \|\cdot\|_F) \le \left\lceil \frac{\|X\|_F^2 b^2}{\epsilon^2} \right\rceil \ln(2dm).$$

Proof.

Let $W \in \mathbb{R}^{m\times d}$ be given with $\|W^\top\|_{2,1} \le r$. Define $s_{ij} := W_{ij}/|W_{ij}|$, and note

$$\begin{aligned} WX^\top &= \sum_{i,j} e_i e_i^\top W e_j e_j^\top X^\top = \sum_{i,j} e_i W_{ij} (Xe_j)^\top \\ &= \sum_{i,j} \frac{|W_{ij}|\|Xe_j\|_2}{r\|X\|_F} \frac{r\|X\|_F s_{ij} e_i (Xe_j)^\top}{\|Xe_j\|} = \sum_{i,j} q_{ij} \times U_{ij}. \end{aligned}$$

Note by Cauchy-Schwarz that

$$\sum_{i,j} q_{ij} \le \frac{1}{r\|X\|_F} \sum_i \sqrt{\sum_j W_{ij}^2} \|X\|_F = \frac{\|W^\top\|_{2,1}\|X\|_F}{r\|X\|_F} \le 1.$$

□

Proof of the Lemma 1

Proof.

potentially with strict inequality, thus q is not a probability vector, which we will want later. To remedy this, construct probability vector p from q by adding in, with equal weight, some U_{ij} and its negation, so that the above summation form of $WX^\top$ goes through equally with p as with q. Now define IID random variables $(V_1, \dots, V_k)$, where

$$\begin{aligned}
&\Pr[V_\ell = U_{ij}] = p_{ij}, \\
&\mathbb{E} V_\ell = \sum_{i,j} p_{ij} U_{ij} = \sum_{i,j} q_{ij} U_{ij} = WX^\top, \\
&\|U_{ij}\| = \left\| \frac{s_{ij} e_i (X e_j)}{\|X e_j\|_2} \right\|_F r \|X\|_F = |s_{ij}| \|e_i\|_2 \left\| \frac{X e_j}{\|X e_j\|_2} \right\|_2 r \|X\|_F = r\|X\|_F, \\
&\mathbb{E} \|V_\ell\|^2 = \sum_{i,j} p_{ij} \|U_{ij}\|^2 \le \sum_{i,j} p_{ij} r^2 \|X\|_F^2 = r^2 \|X\|_F^2.
\end{aligned}$$

$\square$

Proof of the Lemma 1

Proof.

By Lemma 5.1 (Maurey (Pisier 1980)), there exist $(\hat{V}_1, \ldots, \hat{V}_k) \in S^k$ with

$$\left\| WX^\top - \frac{1}{k} \sum_\ell \hat{V}_\ell \right\|^2 \leq \mathbb{E} \left\| \mathbb{E} V_1 - \frac{1}{k} \sum_\ell V_\ell \right\|^2 \leq \frac{1}{k} \sum_\ell \|V_1\|^2 \leq \frac{r^2 \|X\|_F^2}{k}$$

Furthermore, the matrices $\hat{V}_\ell$ have the form

$$\frac{1}{k} \sum_\ell \hat{V}_\ell = \frac{1}{k} \sum_\ell \frac{s_\ell e_{i_\ell} (X e_{j_\ell})^\top}{\|X e_{j_\ell}\|} = \left[\frac{1}{k} \sum_\ell \frac{s_\ell e_{i_\ell} e_{j_\ell}^\top}{\|X e_{j_\ell}\|} \right] X^\top,$$

by this form, there are at most $(2nd)^k$ choices for $(\hat{V}_1, \ldots, \hat{V}_k)$. $\square$

Proof of the Lemma 2

Lemma

Let $\mathcal{F}_n$ be the same image vectors as in the theorem, and let per-layer tolerances $(\epsilon_1, ..., \epsilon_L)$ be given. then

$$\ln \mathcal{N}(\mathcal{F}_n, \sum_{j=1}^{L} \rho_j \epsilon_j \Pi_{k=j+1}^{L} \rho_k s_k, \| \cdot \|_F) \leq \sum_{i=1}^{L} \left\lceil \frac{\|X\|_F^2 b_i^2 \Pi_{j<i} \rho_j^2 s_j^2}{\epsilon_i}^2 \right\rceil \ln(2m^2).$$

Proof.

Let X_i denote the output of layer i of the network, using weights $(W_i, ..., W_1)$, meaning

$$X_0 := X \qquad \text{and} \qquad X_i := \sigma_i(X_{i-1} W_i^\top).$$

The proof recursively constructs cover elements $\hat{X}_i$ and weights $\hat{W}_i$ for each layer with the following basic properties. □

Proof of the Lemma 2

Proof.

- Define $\hat{X}_0 := X_0$, and $\hat{X}_i := \Pi_{B_i} \sigma_i(\hat{X}_{i-1} \hat{W}_i^\top)$, where B_i is the Frobenius-norm ball of radius $\|X\|_F \Pi_{j<i} \rho_j s_j$.
- Due to the projection Π_{B_i}, $\|\hat{X}_i\|_F \leq \|X\|_F \Pi_{j<i} \rho_j s_j$. Similarly, using $\rho_i(0) = 0$, $\|X_i\|_F \leq \|X\|_F \Pi_{j<i} \rho_j s_j$.
- Given $\hat{X}_{i-1}$, choose $\hat{W}_i$ via Lemma above so that $\|\hat{X}_{i-1} W_i^\top - \hat{X}_{i-1} \hat{W}_i^\top\|_F \leq \epsilon_i$, whereby the corresponding covering number $\mathcal{N}_i$ for this layer satisfies

$$\ln \mathcal{N}_i \leq \left\lceil \frac{\|\hat{X}_{i-1}\|_F^2 b_i^2}{\epsilon_i^2} \right\rceil \ln(2m^2) \leq \left\lceil \frac{\|X\|_F^2 b_i^2 \Pi_{j<i} \rho_j^2 s_j^2}{\epsilon_i}^{2} \right\rceil \ln(2m^2).$$

$\square$

Proof of the Lemma 2

Proof.

- Since each cover element $\hat{X}_i$ depends on the full tuple $(\hat{W}_i, \ldots, \hat{W}_1)$, the final cover is the product of the individual covers (and not their union), and the final cover log cardinality is upper bounded by

$$\ln \Pi_{i=1}^{L} \mathcal{N}_i \leq \sum_{i=1}^{L} \left\lceil \frac{\|X\|_F^2 b_i^2 \Pi_{j<i} \rho_j^2 s_j^2}{\epsilon_i}^2 \right\rceil \ln(2m^2).$$

It remains to prove, by induction, an error guarantee

$$\|X_i - \hat{X}_i\|_F \leq \sum_{j=1}^{i} \rho_j \epsilon_j \Pi_{k=j+1}^{i} \rho_k s_k .$$

The base case $\|X_0 - \hat{X}_0\|_F = 0 = \epsilon_0$ holds directly. For the inductive step, by the above ingredients and the triangle inequality,

$\square$

Proof of the Lemma 2

Proof.

$$
\begin{aligned}
\|X_i - \hat{X}_i\|_F &\leq \rho_i \|X_{i-1} W_i^\top - \hat{X}_{i-1} \hat{W}_i^\top\|_F \\
&\leq \rho_i \|X_{i-1} W_i^\top - \hat{X}_{i-1} W_i^\top\|_F + \rho_i \|\hat{X}_{i-1} W_i^\top - \hat{X}_{i-1} \hat{W}_i^\top\|_F \\
&\leq \rho_i s_i \|X_{i-1} - \hat{X}_{i-1}\|_F + \rho_i \epsilon_i \\
&\leq \rho_i s_i [\sum_{j=1}^{i-1} \rho_j \epsilon_j \Pi_{k=j+1}^{i-1} \rho_k s_k] + \rho_i \epsilon_i \\
&= [\sum_{j=1}^{i-1} \rho_j \epsilon_j \Pi_{k=j+1}^{i} \rho_k s_k] + \rho_i \epsilon_i \\
&= \sum_{j=1}^{i} \rho_j \epsilon_j \Pi_{k=j+1}^{i} \rho_k s_k .
\end{aligned}
$$

□

THANK YOU!